Exploratory Data Analysis using ggplot2

Varis A.

Content

Loading libraries

```
library(tidyverse)
library(patchwork)
```

Analysis I. Carbon Dioxide Uptake in Grass Plants

Q : Are plants from different origins have the same cold tolerance?

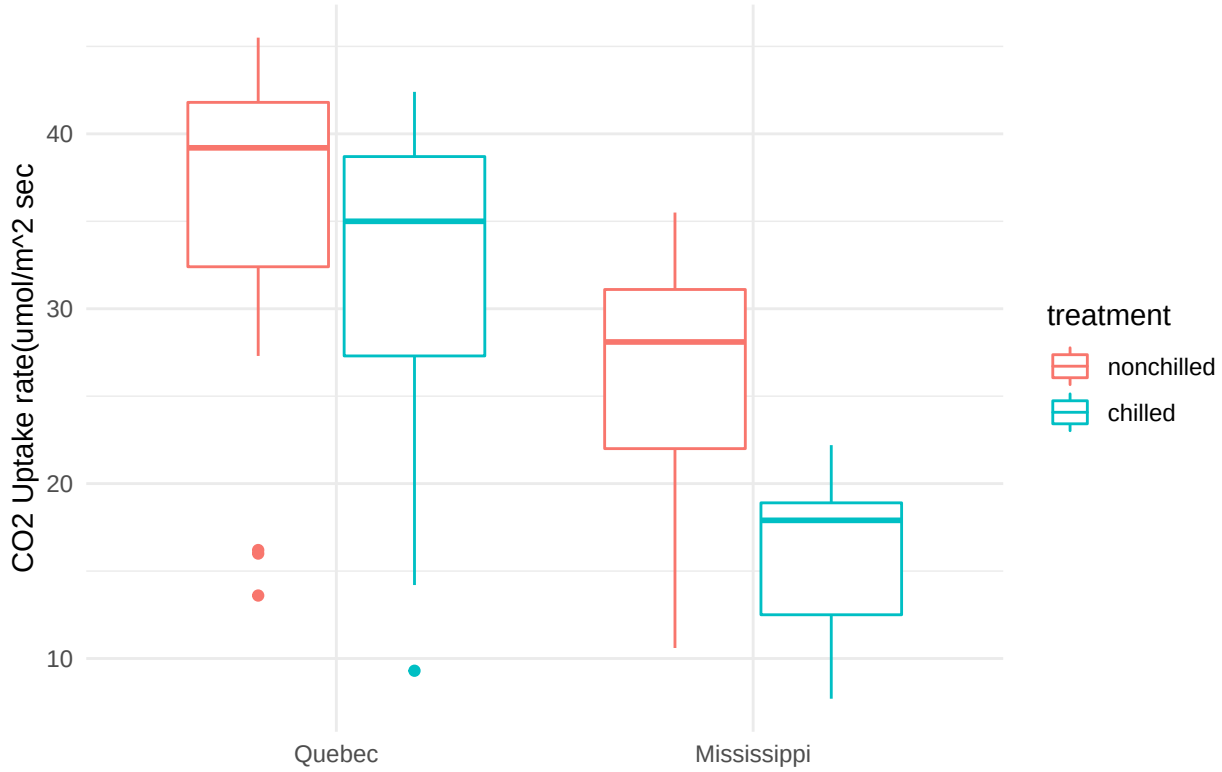
Prepare data

- Change all column names to lowercase for easier operation

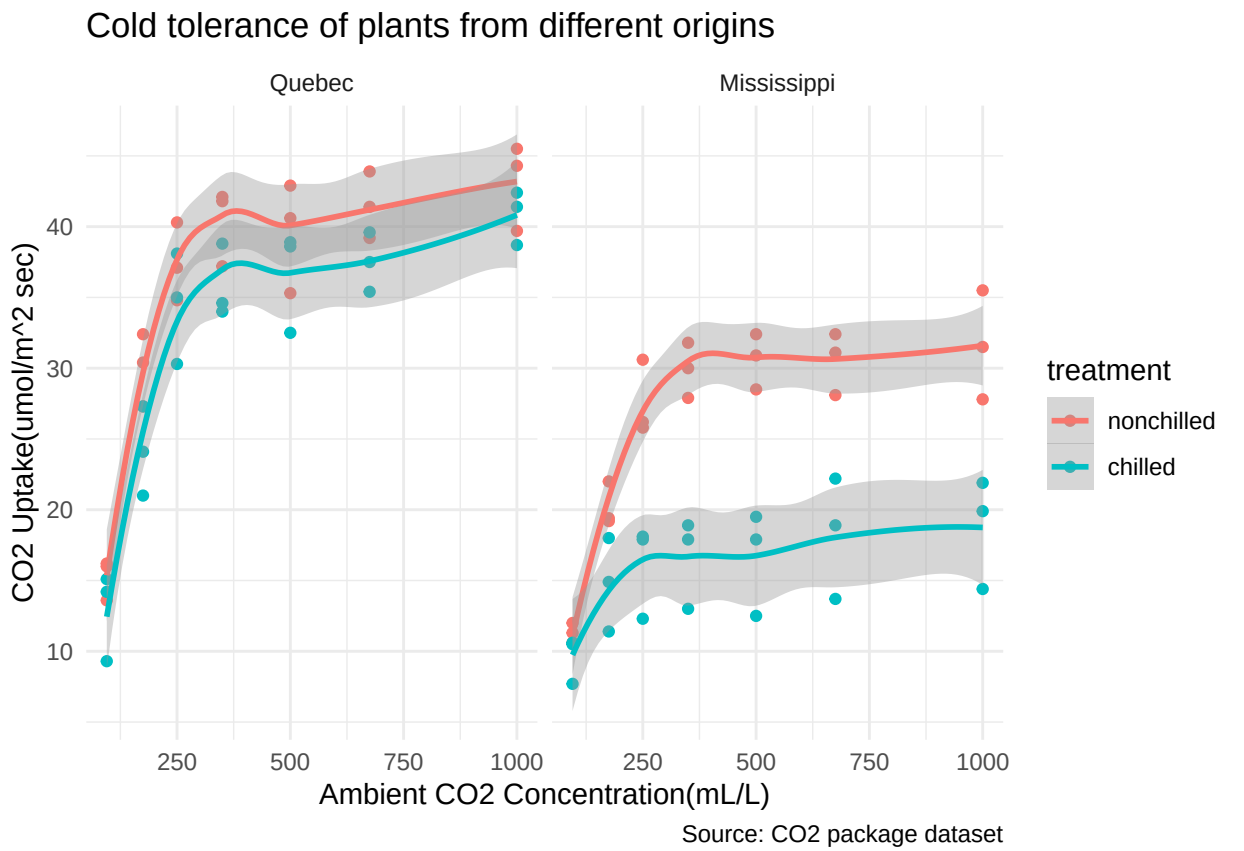
```
co2 <-tibble(CO2) %>%
  mutate(plant=Plant,type=Type,treatment=Treatment) %>%
  select(plant,type,treatment,conc,uptake)
```

Plot graph using `ggplot()`

```
ggplot(co2,aes(type,uptake,color = treatment))+
  geom_boxplot()+
  theme_minimal()+
  labs(x = '',
       y = 'CO2 Uptake rate(umol/m^2 sec',
       caption='Source: CO2 package dataset')
```




```
ggplot(co2,aes(conc,uptake,color=treatment))+
  geom_point()+
  theme_minimal()+
  geom_smooth()+
  facet_wrap(~type)+
  labs(title='Cold tolerance of plants from different origins',
       x = 'Ambient CO2 Concentration(mL/L)',
       y = 'CO2 Uptake(umol/m^2 sec)',
       caption = 'Source: CO2 package dataset')
```



This graph shows that plants from Quebec have more resistant to cold temperature than plants from Mississippi

---

Analysis II. Wine quality based on physicochemical tests

Q : How each factors affect the quality of wine?

Prepare data

- Gather factor data into one columns for easier plotting

```
wine_red <- winequality_red %>%
  gather(c(`citric acid`,chlorides,pH,sulphates,alcohol),key='factor',value='value') %>%
  select(factor,value,quality)
```

- Find the average value of each factor on each quality score to see the trend.

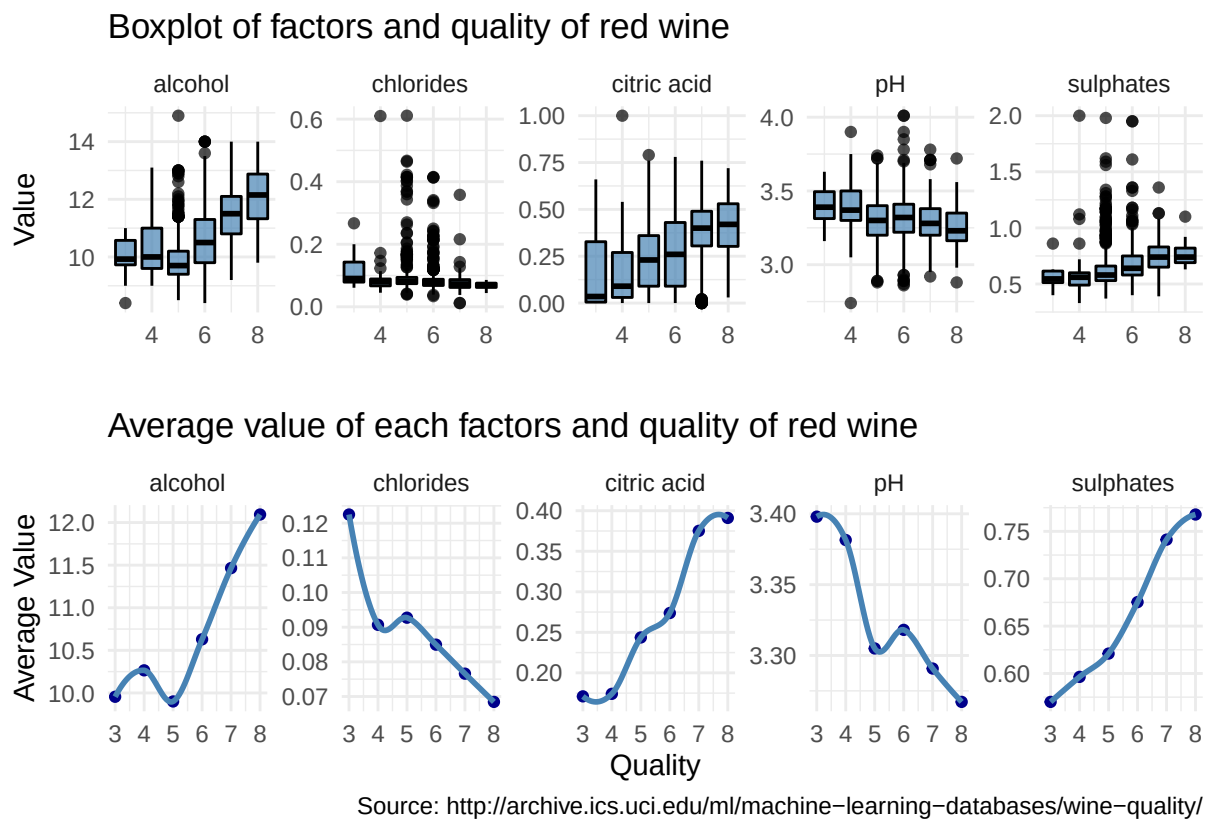
```
sumwine_red <- winequality_red %>%
  group_by(quality) %>%
  summarise(alcohol=mean(alcohol),chlorides=mean(chlorides),'citric acid'=mean(`citric acid`),pH=mean(pH
  gather(alcohol:sulphates,key='factor',value='avg_value')
```

Plot the graphs

```
box_red <- wine_red %>%
  ggplot(aes(quality,value,group=quality))+
  geom_boxplot(color='black',fill='steelblue',alpha=0.7)+
  facet_wrap(~factor,scales='free',nrow=1)+
  theme_minimal()+
  labs(title='Boxplot of factors and quality of red wine',x='',y='Value')

line_red <- sumwine_red %>%
  ggplot(aes(quality,avg_value))+
  geom_point(color='darkblue')+
  geom_smooth(color='steelblue')+
  facet_wrap(~factor,scales = 'free',nrow=1)+
  theme_minimal()+
  labs(title='Average value of each factors and quality of red wine',x='Quality',y='Average Value')

(box_red/line_red)+
  labs(caption="Source: http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/")
```



Now, let's see white wine data

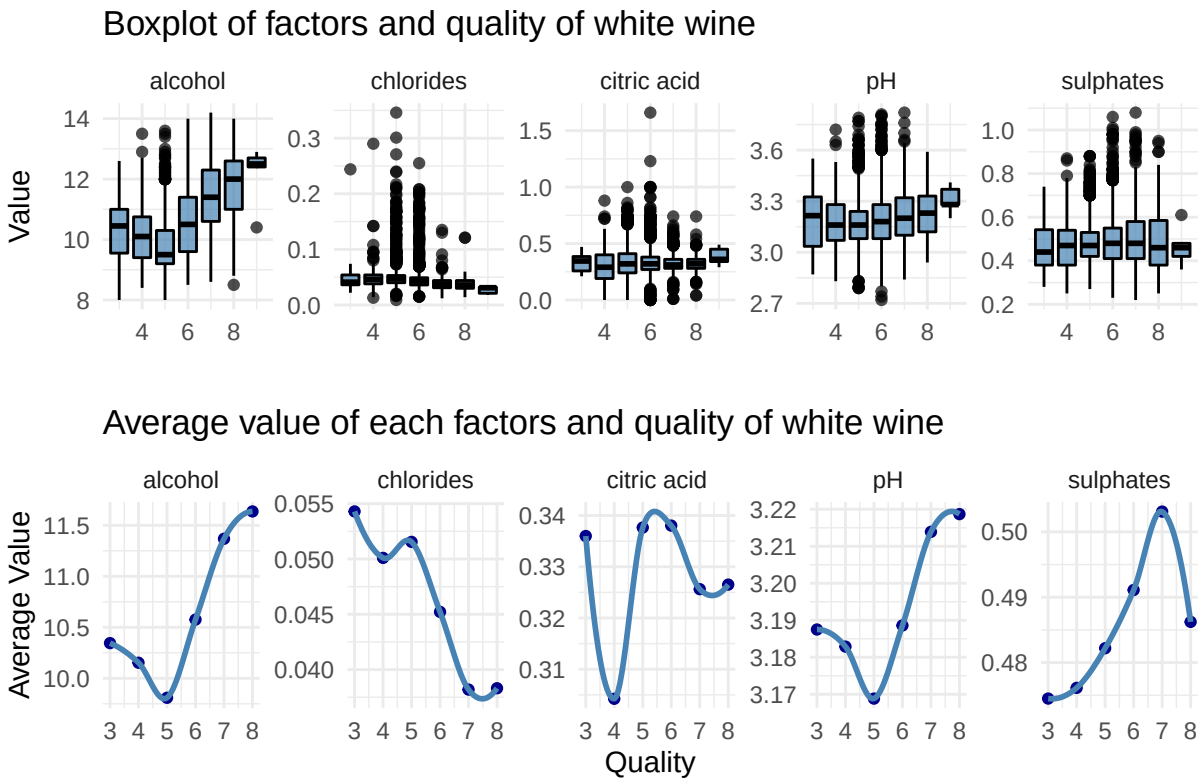
```
wine_white <- winequality_white %>% gather(c(`citric acid`,chlorides,pH,sulphates,alcohol),key='factor'
  select(factor,value,quality)
```

```
sumwine_white <- winequality_white %>%
  filter(quality<9) %>%
  group_by(quality) %>%
  summarise(alcohol=mean(alcohol),chlorides=mean(chlorides),'citric acid'=mean(`citric acid`),pH=mean(pH
  gather(alcohol:sulphates,key=factor,value=avg_value)
```

```
box_white<- wine_white %>%
  ggplot(aes(quality,value,group=quality))+
  geom_boxplot(color='black',fill='steelblue',alpha=0.7)+
  facet_wrap(~factor,scales='free',nrow=1)+
  theme_minimal()+
  labs(title='Boxplot of factors and quality of white wine',x='',y='Value')

line_white <- sumwine_white %>%
  ggplot(aes(quality,avg_value))+
  geom_point(color='darkblue')+
  geom_smooth(color='steelblue')+
  facet_wrap(~factor,scales = 'free',nrow=1)+
  theme_minimal()+
  labs(title='Average value of each factors and quality of white wine',x='Quality',y='Average Value')

(box_white/line_white)+
  labs(x='Quality',caption='Source: http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quali
```

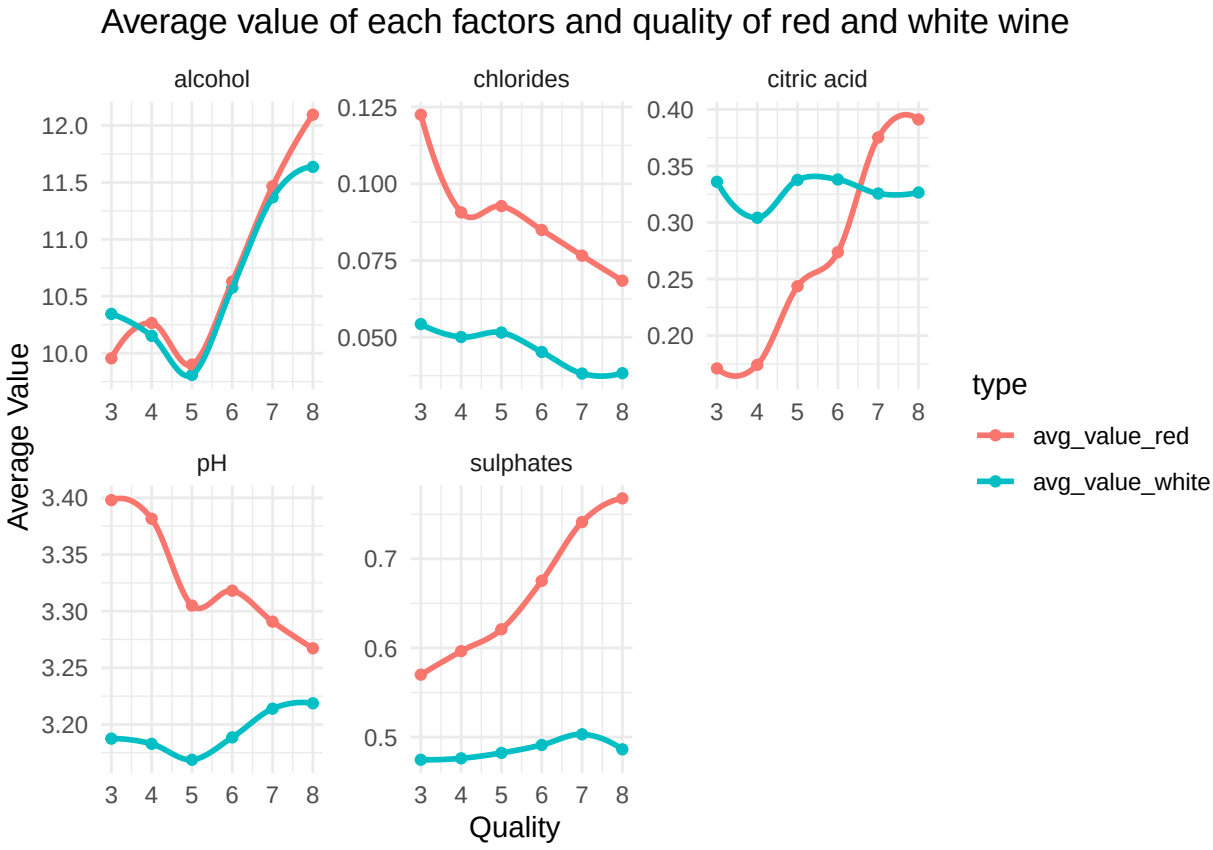


Source: http://archive.ics.uci.edu/ml/machine−learning−databases/wine−quality/

Compare between red and white wine

```
sumwine <- sumwine_white %>%
  inner_join(sumwine_red,by=c('quality','factor')) %>%
  select(quality,
         factor,
         avg_value_white=avg_value.x,
         avg_value_red=avg_value.y) %>%
  gather(avg_value_white:avg_value_red,key= 'type',value='avg_value')

sumwine %>%
  ggplot(aes(quality,avg_value,color=type))+
  geom_point()+
  geom_smooth(se=F)+
  facet_wrap(~factor,scales = 'free',nrow=2)+
  theme_minimal()+
  labs(title='Average value of each factors and quality of red and white wine',x='Quality',y='Average Va
```



Analysis III. Netflix Daily Top 10

Q : What are the top 10 long lasting Netflix original content in Netflix

Prepare data

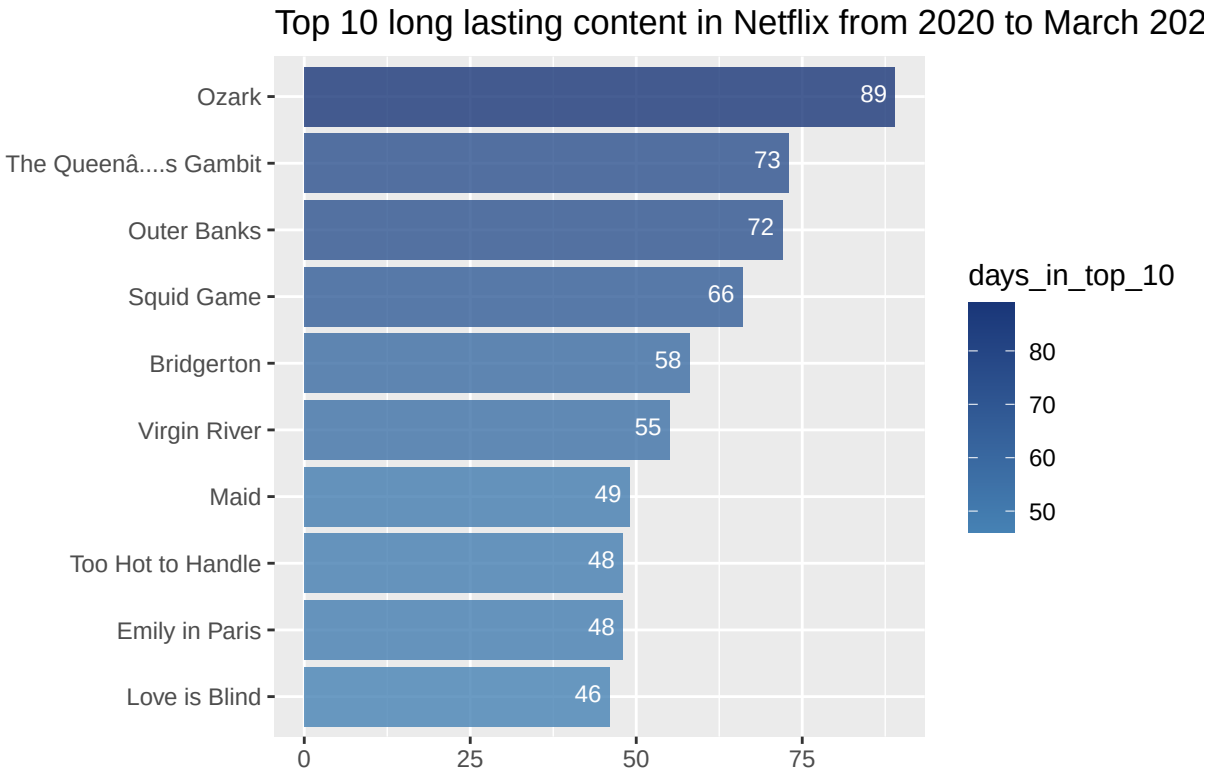
```
netflix <- netflix_daily_top_10 %>%
  select(asof=`As of`,rank=Rank,ytdran=`Year to Date Rank`,
         lastweekrank=`Last Week Rank`,title=Title,type=Type,
         netflixexclusive=`Netflix Exclusive`,
         netflix_release_date=`Netflix Release Date`,
         days_in_top_10=`Days In Top 10`,
         viewership_score=`Viewership Score`)

top10 <- netflix %>%
  filter(netflixexclusive=='Yes') %>%
  group_by(title) %>%
  summarise(days_in_top_10=max(days_in_top_10)) %>%
  arrange(desc(days_in_top_10)) %>%
  head(10) %>%
```

```
  arrange(days_in_top_10) %>%
  mutate(title=factor(title, levels=title))
```

Plot the graph

```
top10 %>%
  ggplot(aes(days_in_top_10,title,fill=days_in_top_10))+
  geom_col(alpha=0.8)+
  geom_text(aes(label=days_in_top_10),
            hjust=1.3,
            vjust=0.3,
            size=3,color='white')+
  labs(title='Top 10 long lasting content in Netflix from 2020 to March 2022,',
       x='',
       y='',caption='Source:https://www.kaggle.com/datasets/prasertk/netflix-daily-top-10-in-us')+
  scale_fill_gradient(low='steelblue',high='#193678')
```



Source:https://www.kaggle.com/datasets/prasertk/netflix-daily-top-10-in-us

Analysis IV. Starbucks Locations

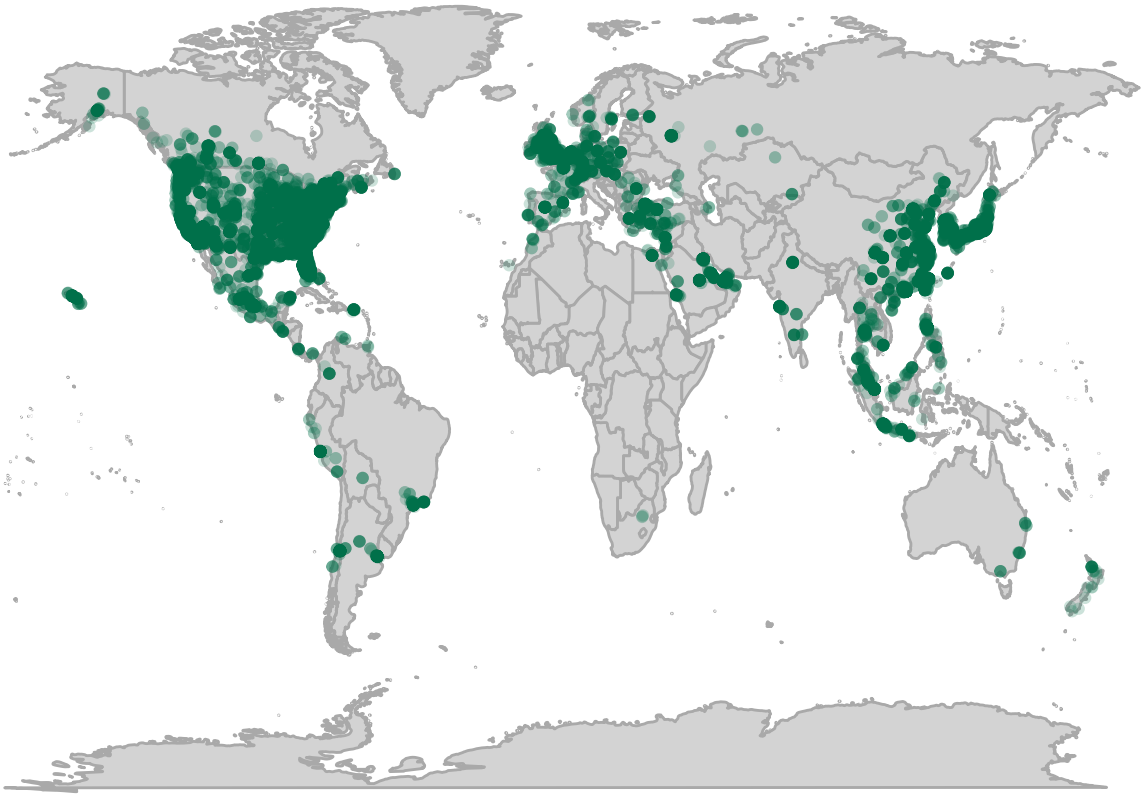
Q : Which country has the most Starbucks stores

Prepare data

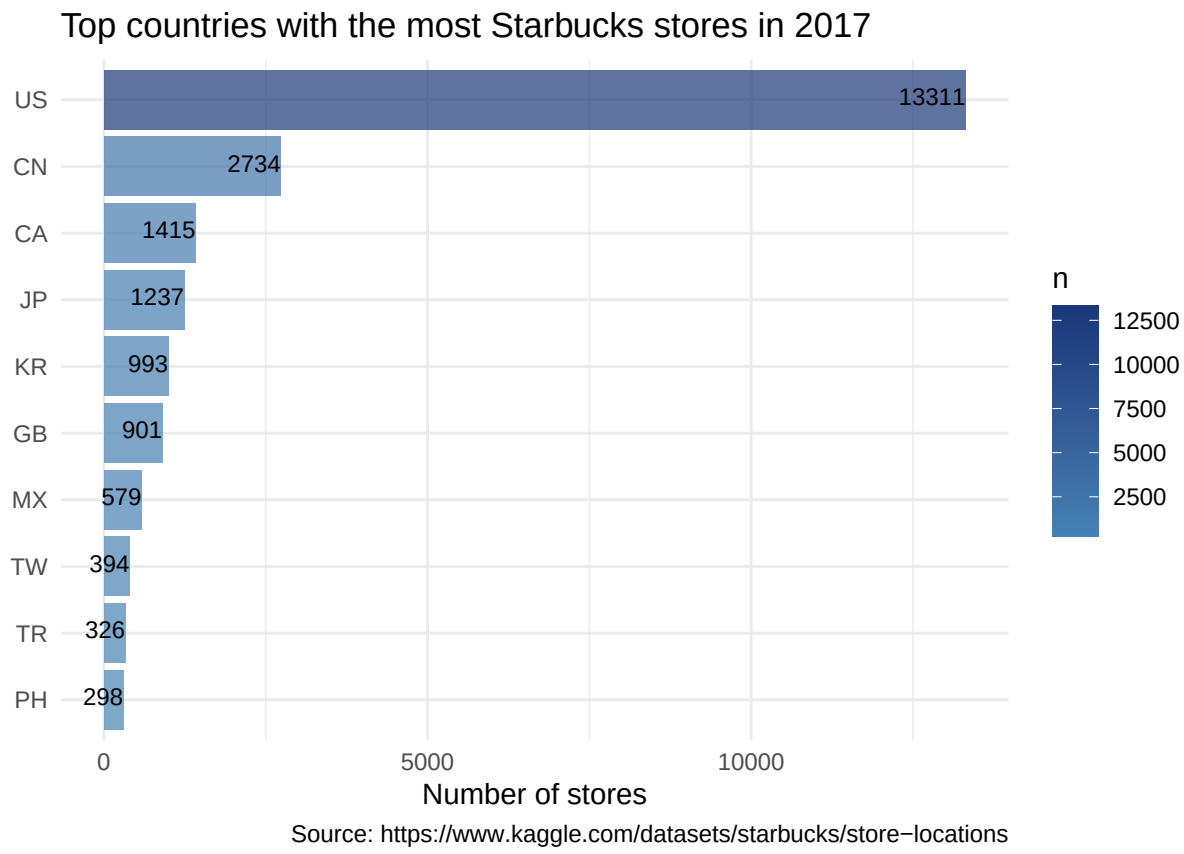
```
directory <- directory %>%
  select(Brand,
         `Ownership Type`,
         City,
         `State/Province`,
         Country,
         Longitude,Latitude) %>%
  filter(Brand=='Starbucks')
```

Where are the stores?

```
world <- map_data("world")
ggplot()+
  geom_map(data = world, map = world,
    aes(long, lat, map_id = region),
    color = "darkgrey",
    fill = "lightgray")+
  geom_point(data = directory,
             aes(Longitude,Latitude),
             color='#00704A',alpha = 0.2)+
  theme_void()+
  theme(legend.position = "none")
```



```
directory %>%
  count(Country) %>%
  arrange(desc(n)) %>%
  head(10) %>%
  arrange(n) %>%
  mutate(Country=factor(Country, levels=Country)) %>%
  ggplot(aes(n, Country, fill = n))+
  geom_col(alpha=0.7)+
  scale_fill_gradient(low = 'steelblue',high ='#193678')+
  geom_text(aes(label=n),
            hjust = 1,
            vjust = .3,
            size = 3, color = 'black',)+
  labs(title ='Top countries with the most Starbucks stores in 2017',
       x ='Number of stores',
       y = '',
       caption = 'Source: https://www.kaggle.com/datasets/starbucks/store-locations')+
  theme_minimal()
```


Top countries with the most Starbucks stores in 2017
US
CN
CA
JP
KR
GB
MX
TW
TR
PH
13311
2734
1415
1237
993
901
579
394
326
298
0
5000
10000
Number of stores
n
12500
10000
7500
5000
2500
Source: https://www.kaggle.com/datasets/starbucks/store−locations